Hertie School

Predicting job creation in Colombian cities with key economic, social, and demographic information

Multivariate and univariate regression approach to job growth prediction

Supervised by

PROF. DR. SIMON MUNZERT

Professor of Data Science and Public Policy | Director, Data Science Lab

ALVARO JOSE GUIJARRO MAY

Master of Data Science for Public Policy

2024

Word count: 7572

HERTIE SCHOOL

Berlin, Germany

If man can predict, almost with certainty, those appearances of which he understands the laws; if, even when the laws are unknown to him, experience of the past enables him to foresee, with considerable probability, future appearances; why should we suppose it a chimerical undertaking to delineate, with some degree of truth, the picture of the future destiny of mankind from the results of its history?

(De Condorcet, 1795)

Para Rita, Alvaro, Camilo y Coco. Valió la pena.

I deeply appreciate Prof. Dr. Simon Munzert's support and guidance throughout my Master's journey at the Hertie School.

Table of Contents

1. Introduction

Public administration and its effects have been evolving in parallel with the societies they affect. At the beginning of the XX century, Max Weber highlighted the necessity of a stable distribution of official labors with the end goal of achieving public objectives in his studies about the rationalization of bureaucracy (Weber, 1977). He also insisted on the importance of stablishing a system of rules that clearly specifies the authorities in charge of carrying out said duties, its corresponding functions, and the coercive methods available to guarantee effective policy management (Weber, 1977). He laid out the foundations on which the measurement of public policies was built upon.

To connect these historical insights with contemporary governance, it's important to define that Colombia is a social State under the rule of law, structured under the category of a Unitary Republic, whose central authority is the President of the Republic. Its territorial organization has four autonomous territorial divisions: departments, districts, municipalities, and indigenous territories. The municipalities are the primary and essential units of the territorial organization and, together with the departments, enjoy autonomy, based on the decisions that the decentralized regime allows their local authorities to make (Asamblea Nacional Constituyente, 1991). Each municipality must offer the public services established by law, develop the necessary infrastructure for local progress, plan the growth of its territory, foster community participation, and improve the social and cultural welfare of its residents.

Mayors are conceived in Colombia as heads of local administration, legal representatives, and first political authority (Congreso de la República de Colombia, 1994) as well as the police authority of the municipality. They are democratically elected for four-year terms, without the possibility of reelection (Asamblea Nacional Constituyente, 1991). By virtue of their role as authorities with constitutional functions, mayors act as economic agents, making decisions on the demand and supply of goods and services within their territories and as joint representatives of the authorities with the governors and presidents, who in turn define the policies for the regulation and promotion of labor in Colombia (Dorado, 2021). They must direct the administrative action of their municipalities, ensuring the provision of public services and the proper functioning of local industrial or commercial enterprises. In order to determine their actions and decisions, they must present plans and programs for economic and social development, consistent with municipal expenditure and investment, collection, and budget plans (Asamblea Nacional Constituyente, 1991).

Government officials have the opportunity to guide their communities positively, or negatively, and affect the lives of their citizens with their public policy priorities, implementations, and executions. Leaders specially have the capacity of affecting economic growth in their countries (Jones & Olken, 2005). The actions taken during their administrative periods have effects on current and future economic, social, and demographic metrics, and this relationship is not exclusive for one community; any society that has this type of government structure has and will experience it (Jones & Olken, 2005). A way of determining how positive or negative the performance of public officials has been on their societies is to evaluate key metrics that track important development factors through time.

Being able to measure their overall performance can help public official to "evaluate, control, budget, motivate, promote, celebrate, learn, and improve"(Behn, 2003).

2. Literature review

Do local factors have an effect on employment growth?

There have been previous attempts to try to quantify the effects social, demographic, economic, or geospatial factors have in local employment growth, like the research conducted by Richard Shearmur and Mario Polèse, in which they analyzed the impact of local and structural factor on employment growth in Canada (Shearmur & Polèse, 2007). By analyzing why employment growth occurs on some regions of the country and not in others, they were able to determine that "local (endogenous) and structural (exogenous) factors retain significant explanatory power"(Shearmur & Polèse, 2007) regarding employment growth. Some of the factors used in this study were education levels, population growth, workers' wages, and geographic locations of Canadian regions.

The International Monetary Fund has pointed out that job growth and creation are within countries and cities top priorities, but the outlook for growth and creation remain as an important concern (International Monetary Fund, 2013). The set of forces that influence growth and job creation in developing countries in recent decades are technological change, demographic changes, poverty rates, GDP, income inequality, and fiscal maneuverability (International Monetary Fund, 2013).

In the Colombian context, researchers have looked into the effects fiscal policies (Gerardo et al., 2014) as well as educational and health reforms (Martínez-Álvarez, 2015) have in employment rates and the economy. The country has been looking into using data for predicting labor market indicators at least since the 2010's. In 2013, the International Labor Organization (ILO) in collaboration with the Ministry of Labor and the National Administrative Department of Statistics published a report in which a work projection model was developed. The "Modelo de Proyección de Empleo"(MPE) is a tool the permits the analysis of labor market changes as a result of intersectoral movements and the dynamics of the local economy (OIT, 2013). The construction of this model relies on data obtained from the behavior of Colombia's GDP, financial indicators like imports and exports, private and government consumption, as well as data related to the characteristics of the local labor market, which is collected and compiled every trimester in the "Gran Encuesta Integrada de Hogares"(OIT, 2013).

In 2014, the Sub-Directorate of Labor Analysis, Monitoring and Prospective ("SAMPL" in Spanish) was tasked with the implementation and management of the MPE system from the ILO, since according to the Decree 4108 of 2011, the SAMPL was created to "Conduct studies, analysis and reporting of information and research on the labor market at national and local level to guide decision-making and the formulation of public policy"(Presidencia de la República de Colombia, 2011). The labor performed by the SAMPL showcases one of the first cases of model creation and usage in order to predict labor market indicators in the country (Ministerio del Trabajo, 2014).

In 2016, the United Nations Development Programme (PNUD in Spanish) launched the "Labor Market: Productivity and Competitiveness for Development" project in Colombia, which focuses on the generation, analysis and visualization of socioeconomic information in order to strengthen decision making of key actors in development, labor inclusion, productivity, and competitiveness in the local, regional, and national context (PNUD, 2016). In it, the PNUD with its local partners evaluate different demographic, social, and economic variables like education levels, unemployment, salaries, age distribution, and company size in different economic sectors of the country in order to understand the state of the current labor market. With this information, an adjustment to the MPE model was realized in partnership with SAMPL in order to improve its prediction accuracy (Ministerio del Trabajo, 2014). Unfortunately, there has been a technical restriction to update the scenarios of the MPE model due to changes in the national accounts by DANE (Ministerio del Trabajo, 2014), so there is room to implement and explore other models.

In general, governments all over the world are performing and operating in ever more complex and interconnected ways, in which being able to generate, store, and analyze data becomes a valuable skill to have. Being able to better understand citizens necessities by analyzing different types of indicators is becoming a common practice in successful administrations. Some successful cases, in the United States for example, of the implementation of data analysis in order to tailor better public policies are the city of New Orleans, that developed a prediction model to calculate the risk of fires in the city before they occur, or the city of Chicago that improved the accuracy in the identification by health inspectors of food establishments where serious noncompliance to health regulation might occur (Goldsmith, 2017).

3. Research Question

Objective: To explore the potential of using historical economic, social, and demographic indicators to aid design public policy decisions made by government officials in Colombia's main cities, with the aim to enhance formal employment metrics. This study seeks to assess the feasibility of such predictions, paving the way for future research that could develop a more sophisticated and practical decision-making tool.

Research question: To what extend could the historical evolution of economic, social, and demographic indicators could be utilized to predict formal employment indicators in Colombia's major cities?

4. Methodology

Data

As of 2019, Colombia had a population of around 49 million people, distributed in 32 states and 1103 municipalities (DANE, 2019). 21 million people, or around 45% of the total population of the country is distributed along its 13 biggest cities and their metropolitan areas (referred to as "A.M" for their Spanish definition of "Area Metropolitana"). These are: Bogota D.C, Medellín A.M., Cali A.M., Barranquilla A.M., Cartagena, Cúcuta A.M., Bucaramanga A.M. Villavicencio, Ibagué, Monteria, Pereira A.M., Manizales A.M, and Pasto. The following map shows the distribution of these cities on the Colombian territory with their official names:

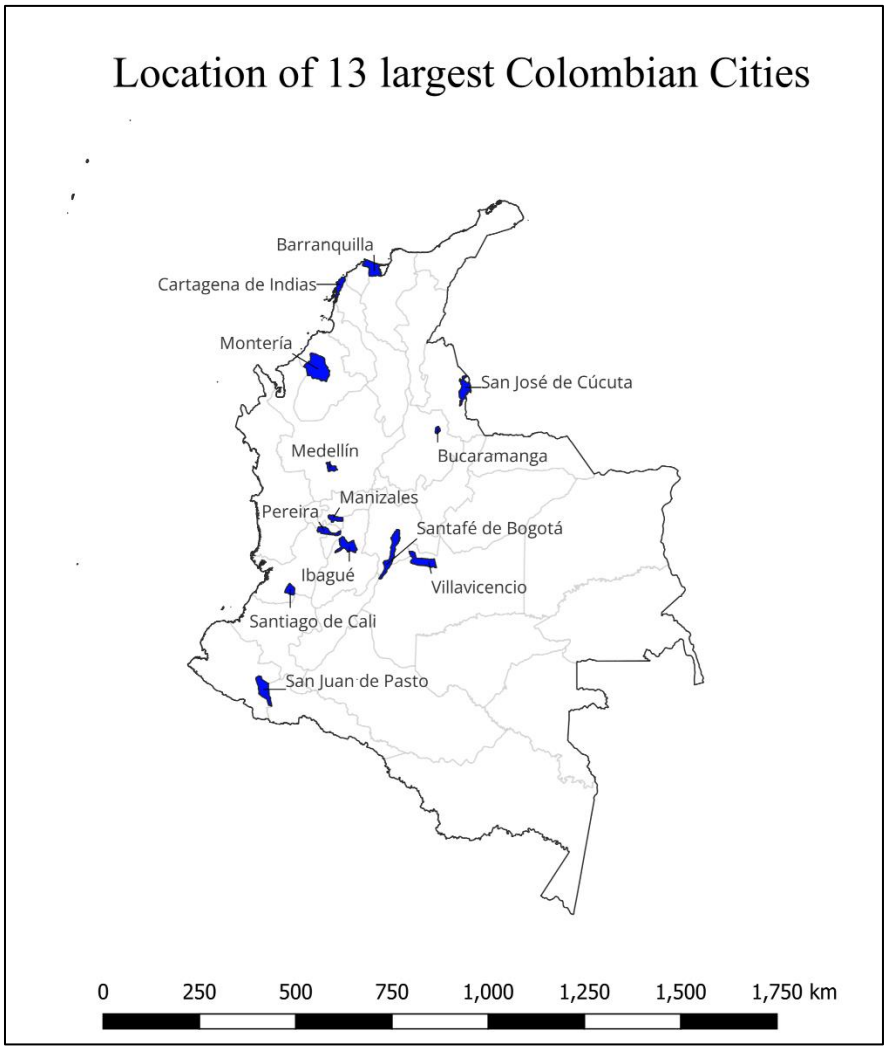


Figure 1 Map of selected Colombian cities

The country's unique socio-economic landscape has changed through the years by various demographic, political, economic, geospatial, and social factors. Most of the population has historically been based in the center of the country, some 2,625m above sea level in the capital city of Bogotá D.C. in the Andes mountains, with other population centers scattered throughout the country's diverse landscape. Here, each city has found their own cultural, economic, and industrial identity, and has managed to provide stability and community for their local populations. Bogotá D.C. and Medellín A.M. are considered the commercial and administrative centers of the country, with Barranquilla A.M. and Cartagena being major port cities filled with trade, industry and tourism, Pereira A.M., Manizales A.M., and Ibagué focusing on coffee and rich agricultural activities, in addition to Cali A.M. and Bucaramanga concentrating on manufacturing and industrial activities. Each city has had its own population evolution during the last decades, and this has also affected their job markets. In the following plot we can observe the evolution of employed citizens vs population growth of each city.

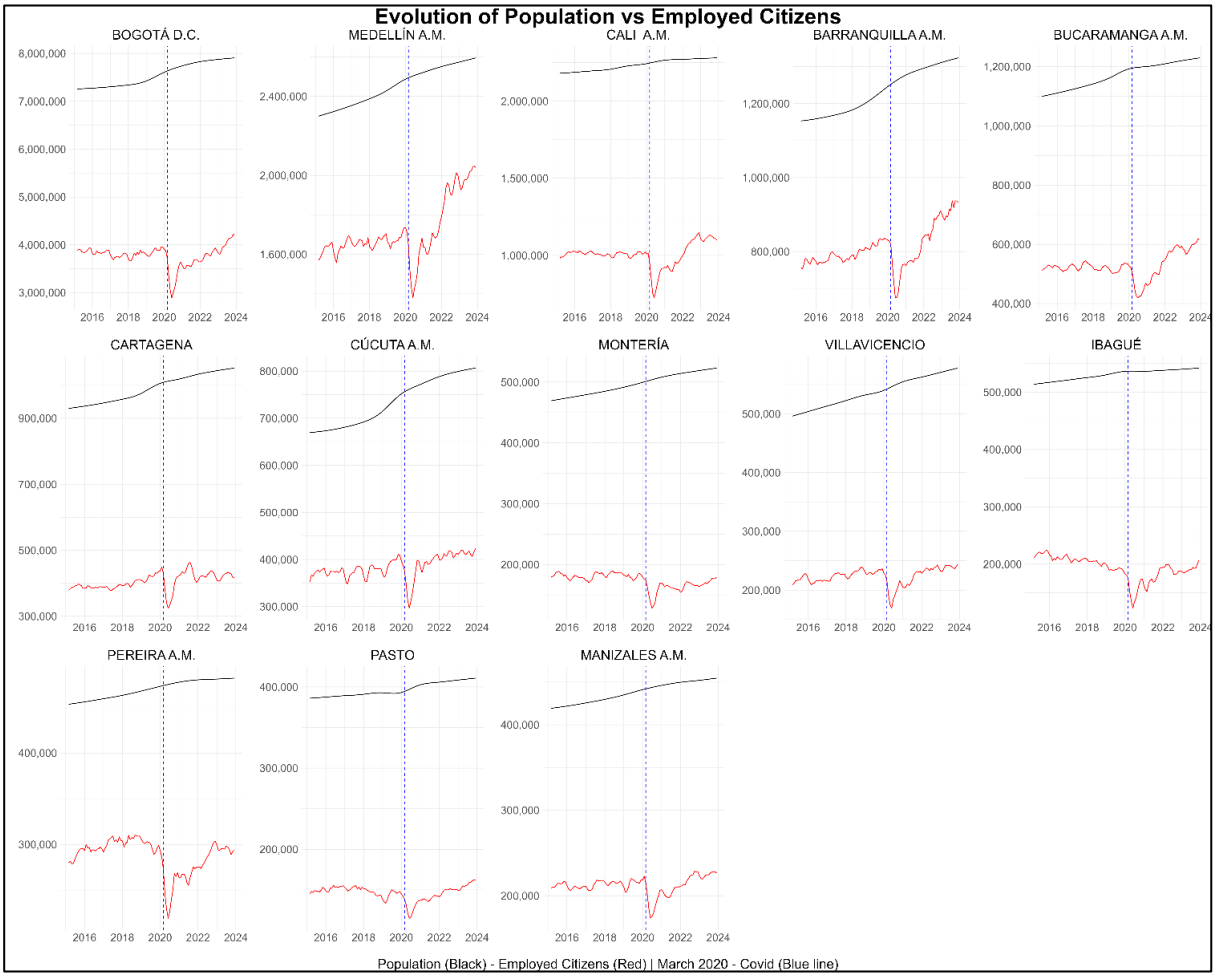


Figure 2 Distribution of population vs employed citizens by city, 2015-2024

All of the cities analyzed in this study suffered from a drop in employed workers after March 2020, which coincides with the beginning of the COVID-19 pandemic in Colombia (Ministerio de Salud y Protección Social, 2020). Most of the cities suffered from a decrease

in their population growth, and since we will be looking into fitting regression models for prediction, the upper time limit chosen for this study is the end of 2019.

Understanding the diverse and unique context of each city is fundamental for attempting to predict their job growth in previous years. Taking inspiration from the literature review, a search for possible datasets was conducted in order to identify the different economic, social, and demographic variables that could be used for predictions. The details of these can be found in Annex 1. In summary, the different datasets gathered for this analysis were:

Name of Dataset	# Of Variables	Source	Description
"*Gran Encuesta Integrada de Hogares*"	4	(DANE, 2023a)	Survey that contains the information of Colombian's employment conditions, in addition to general characteristics of the population such as sex, age, marital status and educational level, and asks about their sources of income. The GEIH provides the country with information at the national level, head, regional, departmental, and for each of the departmental capitals.
Population	4	(DANE, 2019)	Population projections taking as base the 2018 Census methodology.
Consumer Price Index	6	(Banco de la República, 2024)	The consumer price index (CPI) measures the evolution of the average cost of a basket of goods and services representative of households' final consumption, expressed in relation to a base period. Calculated with data from DANE.
Education	24	(Ministerio de Educación Nacional, 2024)	Contains statistical information on preschool, primary, secondary, and high school levels related to sector indicators by municipality without outliers, from 2011 to 2022.
Monetary Poverty	8	(DANE, 2023b)	Contains official monetary poverty figures of the Colombian population, corresponding to the methodological update based on information from the GEIH..
MDM Cities Indicators	24	(Departamento Nacional de Planeación, 2021)	Municipal Performance Measurement (MDM in Spanish) aims to measure, compare, and rank municipalities according to their municipal performance, understood as the management capacity and development results, taking into account their initial status.
Fiscal Performance Amounts	35	(Departamento Nacional de Planeación, 2022a)	Municipal and departmental budget execution information aggregated in their Cash Flow Statement.

Fiscal Performance Scores	10	(Departamento Nacional de Planeación, 2022b)	Fiscal performance Score of the territorial entities for different fiscal years

The information of these datasets and all the data pipeline processes can be accessed in the following GitHub repository: https://github.com/Alvaroguijarro97/Hertie_School_MDS_Master_Thesis

Table 1 Description of datasets gathered

The information contained in each of these datasets could be affected by the public policy decisions of government officials, and following the literature review, have emphasized some influence in job growth and generation indicators (Gerardo et al., 2014; International Monetary Fund, 2013; Martínez-Álvarez, 2015; OIT, 2013; PNUD, 2016; Shearmur & Polèse, 2007). From this initial assessment, a group of 34 variables was selected to create a more manageable dataset for each city. These were:

#	Variables	Description
1	**workers.geih****	**Employed population**
2	date	Year and Month
3	year	Year
4	month	Month
5	population_month.pop	Total Population in monthly frequency (interpolated)
6	population_year.pop	Total Population in yearly frequency
7	CPI.cpi	Consumer Price Index, The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and medical care. The CPI is calculated by taking price changes for each item in the predetermined basket of goods and averaging them.
8	CPI_month_var.cpi	The CPI (Consumer Price Index) monthly variance refers to the change in the CPI from one month to the next, expressed as a percentage. This measure provides an indication of how consumer prices have moved within a month, reflecting short-term inflation or deflation trends.
9	Enrollment_Rate_5_16.edu*	A proportion of the population between 5 and 16 years old are attending the educational system. When DANE's population projections do not adequately capture internal migratory flows, it can reach values greater than 100%."
10	Net_Coverage.edu*	The ratio between the number of students enrolled in transition, primary, secondary, and high school who have the theoretical age (5 to 16 years) and the total population of that same age.
11	Pass_Rate.edu*	Pass rate of students in the official sector. Identifies the percentage of students in preschool, basic, and high school education who pass according to current educational plans and programs.
12	I_PM.mp*	% of population - Monetary Poverty Rate
13	I_PME.mp*	% of population - Extreme Monetary Poverty Rate
14	Gini.mp*	Gini Coefficient (values between 0-1)

15	IPUG.mp*	Average Per Capita Income of the Household Spending Unit in Colombian Pesos
16	LP.mp*	Monetary Poverty Lines (monthly values per person)
17	LPE.mp*	Extreme Monetary Poverty Lines (monthly values per person)
18	MDM_Resource_Mobilization.ci*	Score between 1-100 - Measures mobilization of financial resources
19	MDM_Execution_Of_Resources.ci*	Score between 1-100 - Execution of financial resources
20	MDM_Open_Government_And_Transparency.ci*	Score between 1-100 - Measures of open government and transparency practices
21	MDM_Territorial_Ordering.ci*	Score between 1-100 - Territorial ordering and planning measures
22	MDM_Education.ci*	Score between 1-100 - Educational coverage and quality in middle education
23	MDM_Health_Coverage.ci*	Score between 1-100 - Health coverage and services
24	MDM_Services.ci*	Score 1-100 - Coverage and quality of public services
25	MDM_Security_And_Coexistence.ci*	Score 1-100 - Security and social coexistence indicators
26	TotalIncome.fp*	$ Millions of Pesos - Total income received in the municipality
27	TotalExpenses.fp*	$ Millions of Pesos - Total expenses of the municipality*
28	Self_financing_of_operating_expenses.sfp*	Score 1-100 - Self-financing of operating expenses: the ability to cover the operating expenses of the central administration with unrestricted income (Law 617 of 2000)
29	Debt_service_support.sfp*	Score 1-100 - Debt service support: the ability to support debt service with perceived revenues.
30	Dependence_on_transfers_from_the_Nation_and_Royalties.sfp*	Score 1-100 - Dependence on transfers from the Nation and Royalties: measures the importance of national transfers and royalties (SGR) in total revenues.
31	Generation_of_Own_Resources.sfp*	Score 1-100 - Generation of Own Resources: the ability to generate resources complementary to the transfers.
32	Magnitude_of_Investment.sfp*	Score 1-100 - Magnitude of Investment: quantifies the magnitude of the investment executed by the territorial entity.
33	Saving_Capacity.sfp*	Score 1-100 - Saving Capacity: determines the degree to which surpluses are freed up to finance investment.
34	Fiscal_Performance_Indicator.sfp*	Score 1-100 - Fiscal Performance Indicator

***(interpolated to match GEIH employed workers monthly frequency)**
**** Dependent variable we wish to predict**

Table 2 Selected variables for model fitting and analysis

Preprocessing

The first decision was to define the scope of the research and the timeframe of interest for the evaluation. In Colombia, administrative periods for Mayors and Governors are for 4 years since 2008 (Asamblea Nacional Constituyente, 1991, art 315). Our datasets span between several administrative time periods (2008-2011 / 2012-2015 / 2016-2019 / 2020-2023 / 2024-

2027). Taking into consideration the start of the 2020 Covid-19 pandemic and the disruption it brought to global and local health, economic, logistical, and productive systems, as well as the introduction of the MDM statistic at the end of 2016 ("Medición de Desempeño Municipal" = Municipal Performance Measurement, which ranks Colombian cities by their performance on various key economic, health, safety, and demographic indicators) (Departamento Nacional de Planeación, 2021), the timeframe selected for our predictions analysis was the 2016-2019 administrative time period, precisely from December 2016 until December 2019.

Our variable of interest is the total number of employed civilians for each city, stored in the "Gran Encuesta Integrada de Hogares" (GEIH) dataset. This information is stored in a monthly basis, so in order to be able to fit and define our prediction models, our dependent variables have to match that timescale (Brockwell & Davis, 2010). In order to do so, for all of the variables that were originally stored in a yearly frequency, interpolation was applied. Here, the unknown monthly values were estimated using spline interpolation. In this method of interpolation the interpolant is a piecewise polynomial known as a spline, that fits low-degree polynomials to segments of data points to enhance accuracy and prevent oscillatory errors typical of high-degree polynomial interpolation (Wolber & Alfy, 1999). In order to avoid overestimations in the dependent variables, especially those related to yearly scores, the interpolation was set within unique minimums and maximums for each independent variable, so it followed mostly the trends of their historical information.

Models

In the realm of statistical models, diverse methodologies cater to specific data characteristics and analytical needs. Linear regression models are the foundation of statistical analysis and are used primarily to predict a dependent variable based on linear relationships with one or more independent variables. For this type of model, it is assumed that the relationship between the dependent and independent variables is linear and that the residuals are normally distributed and homoscedastic (Yan & Su, 2009). Time series models are specialized for analyzing data structured in time order, they are crucial for forecasting where data show seasonality, trends, and autocorrelation, and they focus on dependencies within the time series itself rather than external variables (Brockwell & Davis, 2010). Machine learning models provide more robust predictions by learning complex patterns from large datasets without explicitly predefined equations, they can model nonlinear interactions and are particularly useful in scenarios where relationships between variables are highly complex or are not well understood (Hurwitz, 2018). Evaluating these different types of models will provide insights into the feasibility of accurately predicting our variable of interest.

OLS Model

The Ordinary Least Squared (OLS) model is used to estimate the relationship between a dependent variable and one or more independent variables, looking to minimize the sum of the squares of the differences between observed and predicted values. It ensures the best fit line through data points, aiming to yield unbiased and efficient estimates as well as to reduce prediction error (Yan & Su, 2009).

$$Y = \beta_0 + \beta_1 X_i + \cdots + \varepsilon$$

Y = Dependent Variable
β_0 = Intercept of model
β_1= Coefficients of independent variables
X_i = Independent variables
ε = Residuals.

Linear regression makes several key assumptions:

1. There must be a linear relationship between the independent and dependent variables
2. All variables need to be multivariate normal
3. There must be little or no multicollinearity in the data
4. No autocorrelation
5. Homoscedasticity

All of these assumptions need to be met in order to have proper predictions, so the respective tests to check these will be conducted in each of our OLS model pipelines.

The two implemented strategies for fitting and forecasting with OLS models were the following:

1. Comprehensive OLS: Construct an initial OLS model using all predictors, except the date variable, to predict the dependent variable "workers.geih". To refine the model, stepwise regression was employed, which combines backward elimination and forward selection of variables in an iterative process, in order to identify the most significant predictors.
2. Significance-Focused OLS: After the initial stepwise regression, a second OLS model was constructed that incorporated only the predictors identified as statistically significant (with p values less than 0.05), in order to fit a model that empathizes on meaningfulness and interpretability, by focusing only on variables with substantial influence on our predicted outcome.

ARIMA & SARIMA Model

The Autoregressive Integrated Moving Average (ARIMA(*p,d,q*)), model is designed to forecast data based on its own past values and it is a cornerstone of univariate time series analysis (Box & Jenkins, 1976). It encapsulates three key components:

1. Autoregression (AR): Relationship between an observation and a number of lagged observations (*p*)
2. Integration (I): representing the differencing steps to make the series stationary (*d)*
3. Moving Average (MA): models the error term as a combination of previous errors (*q*)

$$\underbrace{\left(1 - \sum_{i=1}^{p} \emptyset_i L^i\right)}_{\text{AR}} \underbrace{(1-L)^d}_{\text{I}} Y_t = \underbrace{(1 + \sum_{j=1}^{q} \theta_j L^j)}_{\text{MA}} \epsilon_t$$

Y_t = Time series data at time *t*, dependent variable we are trying to forecast
p = # of lag observations included in the model
$\emptyset_i$ = Coefficients of the autoregressive terms
L = Lag operator
d = Degree of differencing
q = Order of the moving average
θ_j= Coefficients of the moving average terms
ϵ_t= Error terms

Seasonal Autoregressive Integrated Moving Average (SARIMA) extends the ARIMA model by specifically addressing and modeling seasonal variations in the data. It is particularly useful for modeling data with strong seasonal effects. The model is expressed as SARIMA(*p,d,q*)(*P,D,Q*)[*S*], where (*p,d,q*) have the same definitions as the ARIMA model, *P* represent the seasonal autoregressive, *D* the seasonal differencing, *Q* the moving average terms, and finally *S* indicates the length of the seasonal cycle (Brockwell & Davis, 2010).

$$\underbrace{\left(1-\sum_{i=1}^{p}\emptyset_i L^i\right)}_{\text{Non-seasonal AR}}\underbrace{\left(1-\sum_{I=1}^{p}\Phi_I L^{Is}\right)}_{\text{Seasonal AR}}\underbrace{(1-L)^d}_{\substack{\text{Non-}\\\text{seasonal}\\\text{differencing}}}\underbrace{(1-L^s)^D}_{\substack{\text{Non-}\\\text{seasonal}\\\text{differencing}}}Y_t = \underbrace{(1+\sum_{j=1}^{q}\theta_j L^j)}_{\text{Non-seasonal MA}}\underbrace{(1+\sum_{J=1}^{Q}\Theta_J L^{Js})}_{\text{Seasonal MA}}\epsilon_t$$

Y_t = Time series data at time *t*, dependent variable we are trying to forecast
p = # of lag observations included in the model for non-seasonal autoregressive part
P = # of seasonal autoregressive terms
$\emptyset_i$ = Coefficients of the non-seasonal autoregressive terms
Φ_I = Coefficients of the seasonal autoregressive terms
L = Lag operator
d = Degree of differencing on a non-seasonal level
D = Number of seasonal differencing
q = Order of the non-seasonal moving average
Q = Order of seasonal moving average
θ_j= Coefficients of the non-seasonal moving average terms
Θ_J = Coefficients of seasonal moving average terms
s = length of seasonal cycle in the data
ϵ_t= Error terms

Three types of approaches were handled in this study for the ARIMA and SARIMA models:

1. ARIMA: An initial ARIMA model was automatically fitted by using the auto.arima function in R, where it iterated through the different p, d, q values available and selected the most optimal combination of these.
2. ARIMA with set differencing: As one of the assumption tests for the ARIMA models to check for stationarity, Augmented Dickey-Fuller Tests are applied to our data of

interest to see if differencing is necessary. If it is, the number of differencing rounds applied to the data until it becomes non-stationary is manually set in the auto.arima function in R.

3. SARIMA: A SARIMA model was automatically fitted by using the auto.arima function in R with its seasonal component manually set to TRUE, this will signal the function to consider seasonality in the data of interest and iterated through different p, d, q, P, D, Q values available in order to select the most optimal combination of these (Hyndman & Khandakar, 2008).

Random Forest

Random forest is a machine learning algorithm utilized for predictive modeling, which enhances decision tree methods by generating a "forest" of trees and aggregating their predictions. This technique constructs numerous decision trees during training and determines the outcome by outputting the mean prediction of these trees (Breiman, 2001). Randomness is introduced by using different subsets of the data to build each tree (bootstrap aggregating) and by selecting a random subset of features for each split. It effectively handles high dimensionality and maintains accuracy even with missing values, also known for its robustness against overfitting (Breiman, 2001). It was also selected in this study for its superior capability to model complex interactions and process large datasets efficiently. Three types of approaches were handled for the random forest prediction models:

1. Standard Random Forrest: Baseline model trained on numeric features with a predetermined number of trees (500) and with splits of 10 variables each, without hyperparameter tuning.
2. Optimized Random Forrest for RMSE: Optimized for minimizing the Root Mean Squared Error, with hyperparameter tuning using cross-validation with a random search approach, focusing on selecting the best combination of parameters to lower the RMSE.
3. Optimized Random Forrest for R-Squared: Optimized for maximizing the R-Squared, with hyperparameter tuning using cross-validation with a random search approach, focusing on selecting the best combination of parameters to maximize the R-Squared.

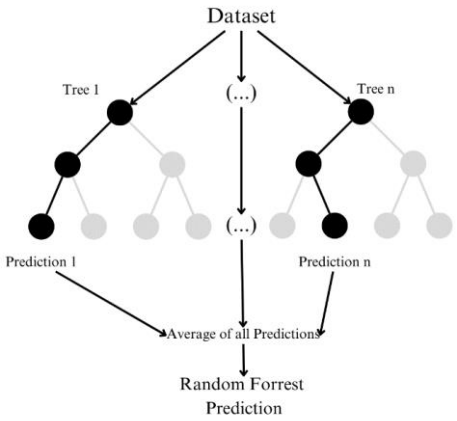


Figure 3 Random Forrest Prediction Algorithm

In order to be able to assess the capability of our selected variables in effectively predicting our employed workers in the different Colombian cities, 8 different models were fitted for each city. For evaluating our models, the following performance metrics were selected for our benchmark:

- **MAE**: Mean Absolute Error; Measures the average magnitude of prediction errors. Lower MAE values indicate a better model fit.
- **RMSE**: Root Mean Squared Error; The square root of the average of the squared differences between prediction and actual observations. Giving high weight to large errors, it is useful when large errors are undesirable, and also lower values are preferred.
- **R-Squared**: Statistical measure of how close the data is to the fitted regression line, in other words the proportion of the variance in the dependent variable that is predicted from the independent variables. ARIMA and SARIMA models don't typically use this metric for evaluation since they focus on minimizing forecast errors and are univariate models (Shumway & Stoffer, 2016).

This iterative approach to modeling and evaluation is imperative in order to determine the best approach to predicting, and to assess the quality and relevance of our models and prediction of our data of interest.

5. Results

Overall

Here we can observe the results for each of our cities, with the model with the lowest MAE to average of employed workers during the timeframe of our prediction (07-2019 / 12-2019) highlighted.

BARRANQUILLA A.M.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	219575	26.35%	302015	97%
OLS (Significant Variables)	39656	4.76%	41950	94%
ARIMA	17775	2.13%	17855	N/A
ARIMA with d=2	**14243**	**1.71%**	**14405**	**N/A**
SARIMA	17775	2.13%	17855	N/A
Random Forest	22257	2.67%	22395	65%
RF Best R_Squared	20121	2.41%	20283	71%
RF Best RMSE	24482	2.94%	24698	58%
Averages	**46986**	**5.64%**	**57682**	**77%**

Table 3. Prediction results for city of BARRANQUILLA A.M.

BOGOTÁ D.C.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	251051	6.41%	284582	95%
OLS (Significant Variables)	115607	2.95%	154838	80%
ARIMA	121960	3.11%	130345	N/A
ARIMA with d=1	**33173**	**0.85%**	**44287**	**N/A**
SARIMA	37204	0.95%	42611	N/A
Random Forest	75592	1.93%	82712	49%
RF Best R_Squared	69052	1.76%	77167	52%
RF Best RMSE	71740	1.83%	78981	54%
Averages	**96922**	**2.47%**	**111941**	**66%**

Table 4. Prediction results for city of BOGOTÁ D.C.

BUCARAMANGA A.M.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	386167	72.75%	599714	98%
OLS (Significant Variables)	158925	29.94%	234134	94%
ARIMA	13616	2.57%	14234	N/A
ARIMA with d=1	21374	4.03%	22395	N/A
SARIMA	15248	2.87%	17783	N/A
Random Forest	16649	3.14%	17645	77%
RF Best R_Squared	19857	3.74%	20704	84%
RF Best RMSE	**11757**	**2.22%**	**12865**	**74%**
Averages	**80449**	**15.16%**	**117434**	**86%**

Table 5. Prediction results for city of BUCARAMANGA A.M.

CALI A.M.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	189383	18.60%	267408	99%
OLS (Significant Variables)	108636	10.67%	145829	98%
ARIMA	11819	1.16%	13099	N/A
ARIMA with d=2	8724	0.86%	9782	N/A
SARIMA	**6794**	**0.67%**	**7037**	**N/A**
Random Forest	14294	1.40%	15490	68%
RF Best R_Squared	15947	1.57%	17043	65%
RF Best RMSE	13316	1.31%	14532	67%
Averages	**46114**	**4.53%**	**61277**	**79%**

Table 6. Prediction results for city of CALI A.M.

CARTAGENA				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	35075	8.22%	39117	98%
OLS (Significant Variables)	11174	2.62%	12721	82%
ARIMA	11467	2.69%	13817	N/A
ARIMA with d=3	80955	18.97%	92605	N/A
SARIMA	**7665**	**1.80%**	**9138**	**N/A**
Random Forest	10737	2.52%	12502	84%
RF Best R_Squared	12313	2.88%	14010	82%
RF Best RMSE	9160	2.15%	11101	84%
Averages	**22318**	**5.23%**	**25626**	**86%**

Table 7. Prediction results for city of CARTAGENA

CÚCUTA A.M.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	104843	26.03%	130793	90%
OLS (Significant Variables)	37726	9.37%	52105	77%
ARIMA	23650	5.87%	26521	N/A
ARIMA with d=3	**10936**	**2.71%**	**16596**	**N/A**
SARIMA	15797	3.92%	17463	N/A
Random Forest	17314	4.30%	18506	64%
RF Best R_Squared	17174	4.26%	18183	71%
RF Best RMSE	15495	3.85%	16997	70%
Averages	**30367**	**7.54%**	**37145**	**74%**

Table 8. Prediction results for city of CÚCUTA A.M.

IBAGUÉ				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	513	0.27%	612	90%
OLS (Significant Variables)	**285**	**0.15%**	**378**	**96%**
ARIMA	1562	0.82%	1919	N/A
ARIMA with d=1	3516	1.84%	4005	N/A
SARIMA	1562	0.82%	1919	N/A
Random Forest	1920	1.00%	2562	83%
RF Best R_Squared	4459	2.33%	5115	80%
RF Best RMSE	1494	0.78%	2188	85%
Averages	**1914**	**1.00%**	**2337**	**87%**

Table 9. Prediction results for city of IBAGUÉ

MANIZALES A.M.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	72773	33.57%	103806	95%
OLS (Significant Variables)	23304	10.75%	36353	94%
ARIMA	2962	1.37%	3807	N/A
ARIMA with d=2	15314	7.06%	16610	N/A
SARIMA	3176	1.47%	3928	N/A
Random Forest	2040	0.94%	2282	68%
RF Best R_Squared	**1566**	**0.72%**	**1961**	**62%**
RF Best RMSE	2233	1.03%	2611	66%
Averages	**15421**	**7.11%**	**21420**	**77%**

Table 10. Prediction results for city of MANIZALES A.M.

MEDELLÍN A.M.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	342053	20.21%	429243	96%
OLS (Significant Variables)	**6518**	**0.39%**	**8033**	**69%**
ARIMA	31189	1.84%	39308	N/A
ARIMA with d=1	30743	1.82%	38902	N/A
SARIMA	16692	0.99%	19680	N/A
Random Forest	25966	1.53%	33853	48%
RF Best R_Squared	32311	1.91%	39963	42%
RF Best RMSE	23731	1.40%	31303	53%
Averages	**63650**	**3.76%**	**80036**	**62%**

Table 11. Prediction results for city of MEDELLÍN A.M.

MONTERÍA				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	22085	12.27%	27548	86%
OLS (Significant Variables)	22085	12.27%	27548	86%
ARIMA	3070	1.71%	3223	N/A
ARIMA with d=2	5244	2.91%	6133	N/A
SARIMA	4437	2.46%	5112	N/A
Random Forest	4431	2.46%	4605	82%
RF Best R_Squared	**4385**	**2.44%**	**4572**	**84%**
RF Best RMSE	4615	2.56%	4814	83%
Averages	**8794**	**4.89%**	**10444**	**84%**

Table 12. Prediction results for city of MONTERÍA

PASTO				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	187594	127.01%	298532	99%
OLS (Significant Variables)	113226	76.66%	167016	98%
ARIMA	4799	3.25%	4969	N/A
ARIMA with d=2	14313	9.69%	16445	N/A
SARIMA	**1203**	**0.81%**	**1537**	**N/A**
Random Forest	1967	1.33%	2353	82%
RF Best R_Squared	2089	1.41%	2565	82%
RF Best RMSE	1697	1.15%	2098	82%
Averages	**40861**	**27.66%**	**61939**	**89%**

Table 13. Prediction results for city of PASTO

PEREIRA A.M.				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	1150147	389.91%	1931806	92%
OLS (Significant Variables)	1169136	396.35%	1959175	90%
ARIMA	6939	2.35%	8026	N/A
ARIMA with d=2	**6854**	**2.32%**	**7937**	**N/A**
SARIMA	7627	2.59%	8654	N/A
Random Forest	8023	2.72%	8906	66%
RF Best R_Squared	7664	2.60%	8651	69%
RF Best RMSE	8135	2.76%	9006	65%
Averages	**295566**	**100.20%**	**492770**	**76%**

Table 14. Prediction results for city of PEREIRA A.M.

VILLAVICENCIO				
Models	**MAE**	**MAE / Avg Employed Workers Of Prediction Time**	**RMSE**	**R_Squared**
OLS	12414	5.33%	17147	96%
OLS (Significant Variables)	**3053**	**1.31%**	**3516**	**40%**
ARIMA	7199	3.09%	7662	N/A
ARIMA with d=3	29363	12.60%	34014	N/A
SARIMA	5384	2.31%	5866	N/A
Random Forest	6525	2.80%	7230	89%
RF Best R_Squared	5819	2.50%	6459	87%
RF Best RMSE	6848	2.94%	7570	88%
Averages	**9576**	**4.11%**	**11183**	**80%**

Table 15. Prediction results for city of VILLAVICENCIO

As we can see, the different models were able to fit the training data and predict employed citizens for each city during the timeframe of our study. There doesn't seem to be an overall optimal model type for all cities, which was expected due to the variety in job growth trends and behavior for each. However, it is important to also visualize these predictions in order to better understand the results. The plots for all of the cities are included in the Annex of this document, but let's analyze the results for BOGOTÁ D.C., BARRANQUILLA A.M., and MEDELLÍN A.M..

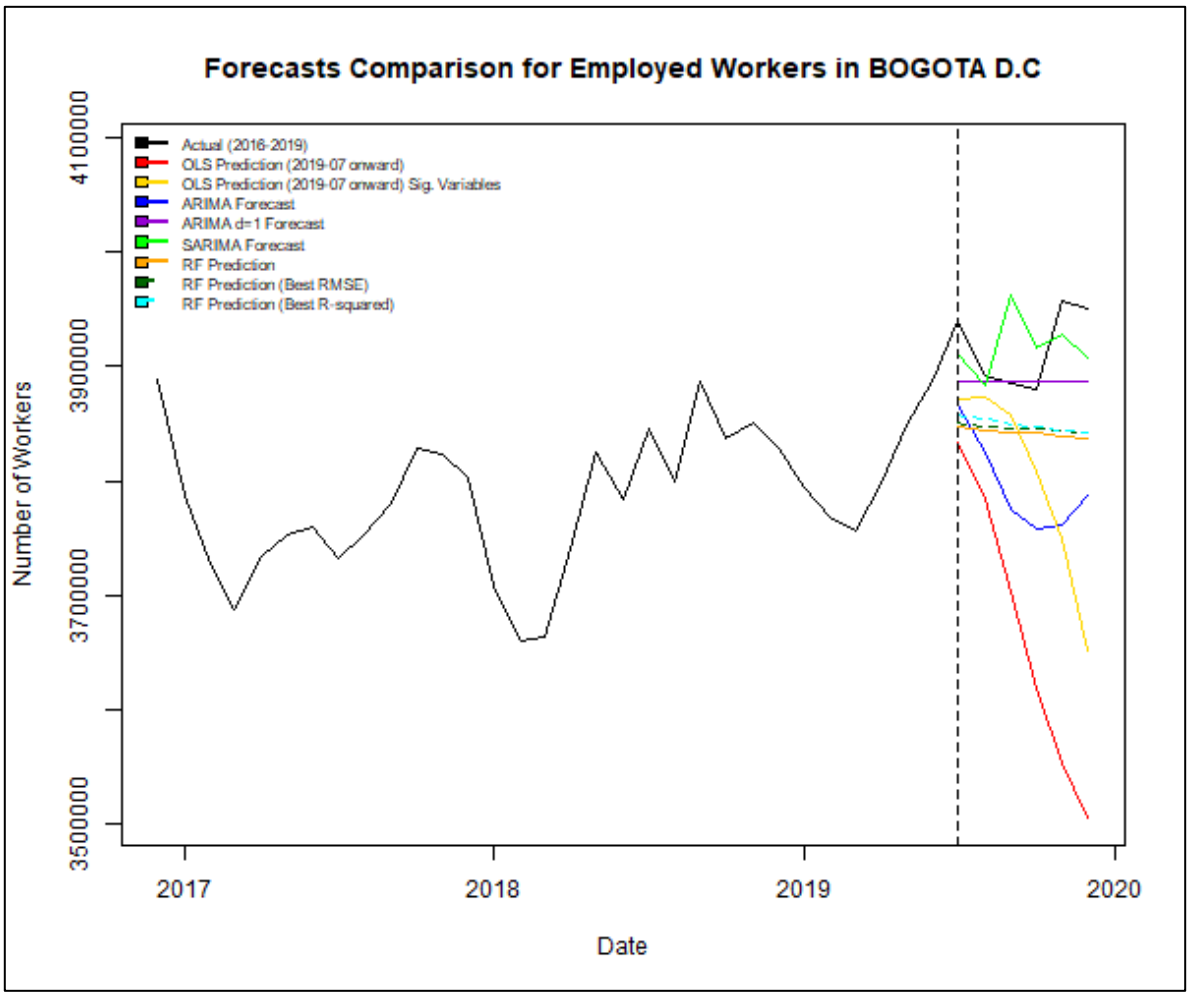


Figure 4 Prediction results for different model types for city of BOGOTA D.C

For the capital city of BOGOTÁ D.C, the model that resulted on the smallest MAE was the ARIMA with differencing = 1, with a value of **33173,** and even though this result is only **0.85%** off the average # of employed citizens during the same timeframe of our prediction (3917496), we can see that the model is not properly picking up the seasonal component of the data and is only predicting a straight line. From the other univariate model types, the SARIMA prediction seems to be better at picking up the past behavior of our data of interest. However, we are more interested in the multivariate models that have been fitted with our variables of interest. None of them seem to be picking up the behavior of the training data correctly and are mostly predicting a straight line. The OLS approaches had the best overall R-Squared from the training data (95% and 80% respectively), but as we can see from their predictions, they are having trouble with handling new data. In average, for the city of BOGOTA D.C, the MAE of all models was 96922, which is 2.47% or the average of employed citizens during the prediction timeframe.

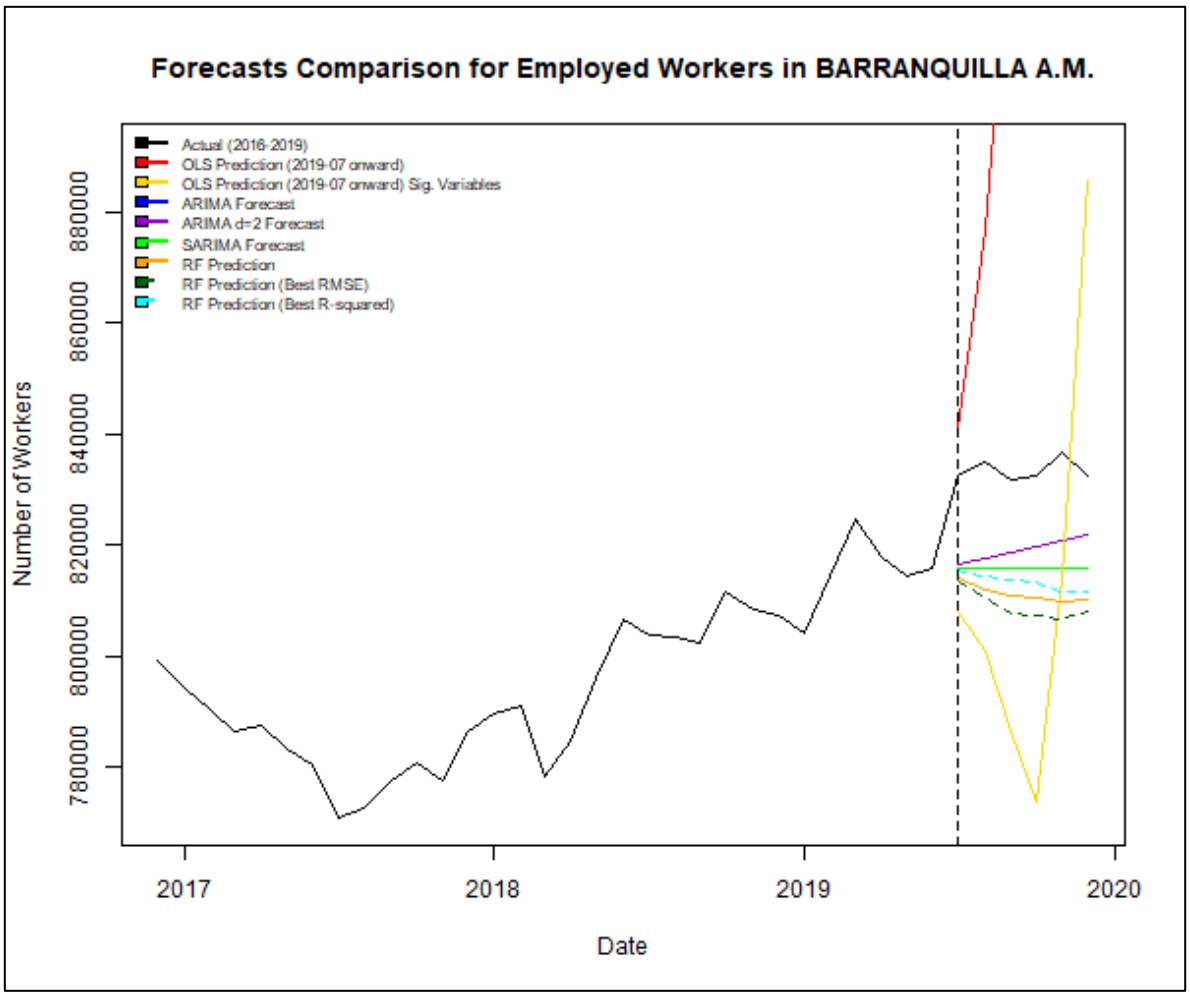


Figure 5 Prediction results for different model types for city of BARRANQUILLA A.M

For the port city of BARRANQUILLA A.M., the model that resulted on the smallest MAE was the ARIMA with differencing = 2, with a value of **14243,** and even though this result is only **1.71%** off the average # of employed citizens during the same timeframe of our prediction (833456), we can see that the model is not properly picking up the seasonal component of the data and is only predicting a straight line following the upward trend of the training data. From the other univariate model types, the SARIMA prediction in this case doesn't seem to be doing a better job at picking up the past behavior of our data of interest. However, we are more interested in the multivariate models that have been fitted with our variables of interest. None of them seem to be picking up the behavior of the training data correctly, with the OLS models significantly deviating from the training data, which might suggest that more robust models are a better fit for the type of data observed. Finally, even though the random forest approaches are more conservative in their prediction, they would still need more refined tunning in order to better pick up the seasonality and overall trend of our data. In average, for the city of BARRANQUILLA A.M., the MAE of all models was 46986, which is 5.64% of the average of employed citizens during the prediction timeframe.

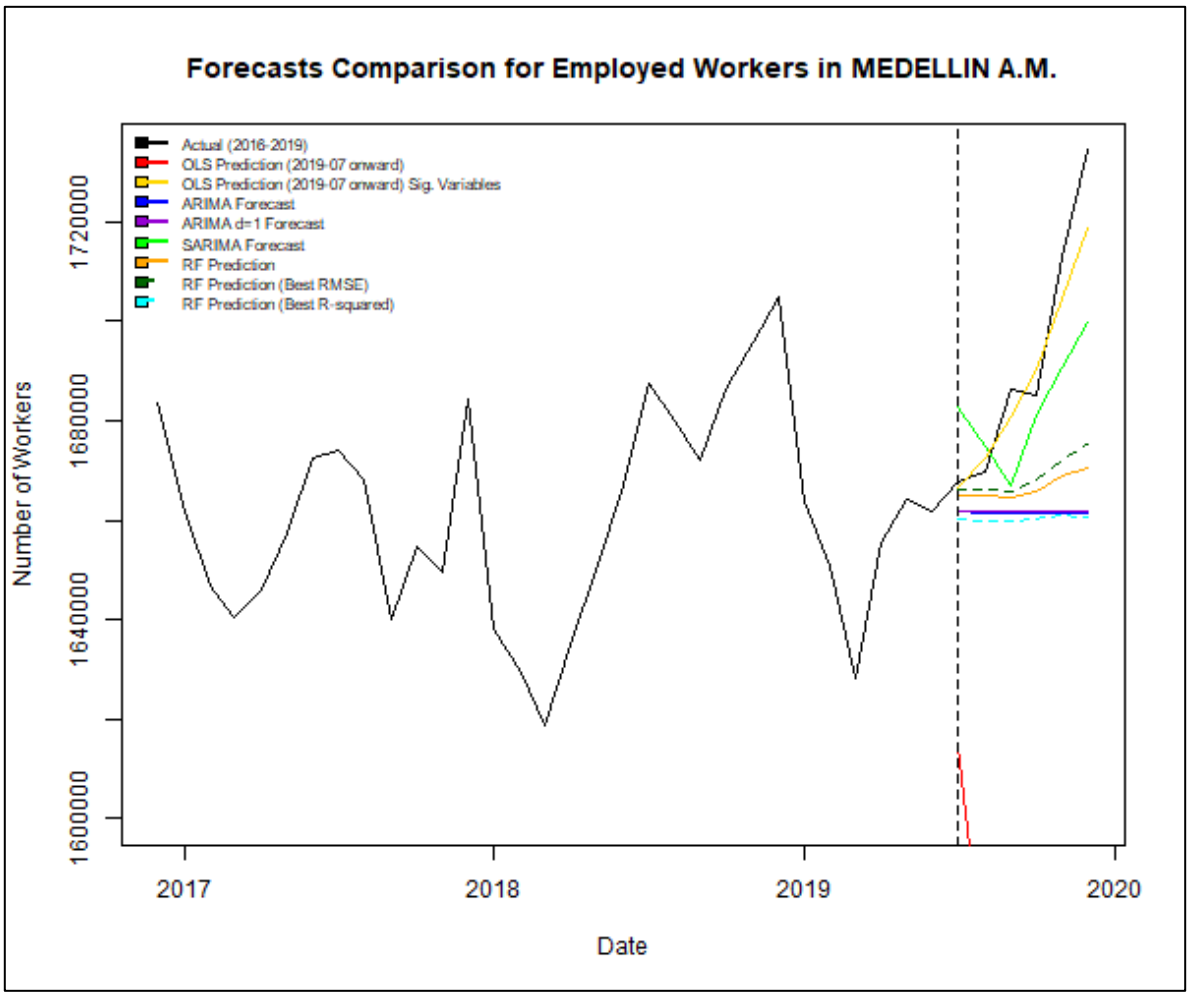


Figure 6 Prediction results for different model types for city of MEDELLIN A.M

For the administrative city of MEDELLIN A.M., the model that resulted on the smallest MAE was the OLS fitted with the most significant values, with a value of **6818,** and even though this result is only **0.39%** off the average # of employed citizens during the same timeframe of our prediction (1692640), we can see that the model is not properly picking up the seasonal component of the data and is only predicting a straight line following the upward trend of the training data. From the other univariate model types, the SARIMA prediction in this case doesn't seem to be doing a better job at picking up the past behavior of our data of interest, and the other ARIMA models are just predicting a straight line. For the multivariate models, in this case having only the significant variables as part of our OLS model gave better results than having the ones selected by the initial stepwise approach, so refinement of the model is preferred, as expected. Finally, even though the random forest approaches are more conservative in their prediction as well, they would still need more refined tunning in order to better pick up the seasonality and overall trend of our data. In average, for the city of MEDELLIN A.M., the MAE of all models was 63650, which is 3.76% of the average of employed citizens during the prediction timeframe.

% Of Cities in which variable was part of the model / top 10 most significant						
Variable Name	OLS	OLS (Significant Variables)	Random Forest Count	RF Best RMSE	RF Best R-Squared	Total
year	69%	62%	0%	0%	23%	31%
month	62%	38%	77%	85%	54%	**63%**
population_month.pop	69%	46%	31%	23%	62%	**46%**
population_year.pop	69%	54%	0%	0%	23%	29%
CPI.cpi	38%	15%	15%	38%	31%	28%
CPI_month_var.cpi	38%	31%	62%	77%	38%	**49%**
Enrollment_Rate_5_16.edu	46%	31%	23%	23%	15%	28%
Net_Coverage.edu	92%	46%	31%	38%	23%	**46%**
Pass_Rate.edu	69%	38%	46%	62%	31%	**49%**
I_PM.mp	69%	62%	15%	23%	31%	**40%**
I_PME.mp	77%	46%	15%	15%	31%	**37%**
Gini.mp	54%	38%	31%	8%	31%	32%
IPUG.mp	77%	54%	8%	8%	54%	**40%**
LP.mp	23%	23%	38%	15%	31%	26%
LPE.mp	23%	8%	54%	54%	23%	32%
MDM_Resource_Mobilization.ci	46%	46%	8%	15%	23%	28%
MDM_Execution_Of_Resources.ci	31%	31%	46%	23%	23%	31%
MDM_Open_Government_And_Transparency.ci	31%	8%	23%	15%	15%	18%
MDM_Territorial_Ordering.ci	69%	8%	31%	23%	23%	31%
MDM_Education.ci	54%	54%	46%	46%	31%	**46%**
MDM_Health_Coverage.ci	69%	54%	15%	38%	15%	**38%**
MDM_Services.ci	31%	23%	46%	38%	31%	34%
MDM_Security_And_Coexistence.ci	8%	8%	38%	31%	23%	22%
TotalIncome.fp	0%	0%	38%	15%	31%	17%
TotalExpenses.fp	0%	0%	31%	31%	54%	23%
Self_financing_of_operating_expenses.sfp	15%	8%	31%	31%	31%	23%
Debt_service_support.sfp	69%	38%	15%	23%	38%	**37%**
Dependence_on_transfers_from_the_Nation_and_Royalties.sfp	23%	23%	31%	38%	31%	29%
Generation_of_Own_Resources.sfp	15%	15%	31%	38%	62%	32%
Magnitude_of_Investment.sfp	23%	23%	23%	54%	31%	31%
Saving_Capacity.sfp	0%	0%	62%	46%	62%	34%
Fiscal_Performance_Indicator.sfp	23%	8%	38%	38%	46%	31%

Table 16. Distribution of importance of selected variables in multivariate models analyzed

In Table 16 we can observe the distribution of our selected variables in the multivariate models for our 13 different city model pipelines. Only one variable was part of the fitted models or belonged to the top 10 most significant variables in the random forest models in more than 50% of the cases, and that was the "month" variable, which might indicate that the seasonality component of our dependent data is a major factor to keep an eye in. The other 9 most repeated variables are highlighted in the "Total" column. Variables that were expected to have a bigger influence in the prediction of job growth in Colombian cities, like those from the Municipal Performance Measurement (MDM) dataset, were not as present in the model selection, with only the Education and Health Coverage scores being present almost 40% of the time.

6. Discussion

Limitations

Modelling and predicting complex macroeconomic variables is an inherently challenging task. There are many external and unknown variables that might affect their behavior and defining what has an effect on their conduct requires a deeper understanding of the subject matter. Simpler models are a good starting point to test the practicality and feasibility of these type of predictions, but in order to obtain more robust and useful models for public policy decision making requires more complex and advanced models to be considered (Ascher, 1981).

The data collected and used for this analysis also plays a major role in determining the quality and usefulness of the predictions. For the time being, most of the data that was collected had a different time frequency from our variable of interest, which meant that some key independent variables had to be interpolated in order to bring them to a useful format. This of course is a major limitation for this analysis, since the quality of the data does not perfectly reflect the real behavior and reality of the economic, demographic, and social variables analyzed. Then again this issue will persist for the close future, since most of the variables of our interest, by the nature of their inception, usefulness, and labor & economic costs for local and national government organizations, will maintain their yearly timescale. Currently, cities and government agencies in Colombia don't have the capacity or the incentives to measure and calculate these variables in a more regular frequency. The National Administrative Department of Statistics (DANE) of Colombia does have some of the key variables used in this study in the granularity and frequency needed for a more thorough modeling, but I was not able to get them in time due to administrative hurdles and delays.

For this initial analysis, scaling and transformation of key economic, social, and demographic variables were not considered since the main objective of this research was to check the plausibility of prediction with these types of variables on job creation metrics. Having different scales and ranges for our independent variables might have been problematic for our modeling pipeline, and correctly handling these aspects of our variables could enhance our modeling and prediction.

Another key limitation of this study was that the scope selected for it was more of a descriptive type, with the causality component between variables being disregarded for the time being. Even though models were fitted, and our interest variable was able to be predicted, further research has to be done in order to understand the relationships and effects our independent variables have on our predictions.

Future Work

Data preprocessing played an important role in this analysis, and in the future different approaches should be explored to enhance the robustness and usefulness of our data. Additionally, looking into the impact of different scaling and transformation methods like logarithmic scaling, normalization, or scaling to a range could further refine our understanding of the data and result in better predictions.

While the models employed in this study have provided a valuable insight, future research could benefit from exploring alternative models that can better capture the complexities of macroeconomic variable predictions, as well as the relationship between demographic, social, and economic indicators. Exploring more powerful models such as LSTM, CNN, or SARIMAX may offer a deeper understanding of the relationships between the data.

This type of modeling approach would also benefit from analyzing prediction models on all the different cities evaluated at the same time, and not treating each city as an individual actor. By taking into account the diversity and unique economic, social, and demographic characteristics of each city in the modeling process, a fairer and more representative prediction of the overall Colombian scenario could be achieved. While considering multiple cities collectively, it is essential to recognize the significance of city-specific variables in shaping economic outcomes as well. Future research could explore methodologies for seamlessly integrating specific city variables into predictive models too.

7. Conclusion

It has been established that prediction of formal employment indicators in Colombian cities is complex, but feasible. Through the implementation of various univariate and multivariate statistical models, this study has demonstrated that even though the number of employed citizens can be forecasted, the accuracy and effectiveness of these predictions vary significantly across different cities and model types. While our forecasts successfully predicted the variable of interest using correlated variables, this does not establish causality, necessitating further investigation into the causal relationships of these variables in order to use these types of model as public policy decision making instruments.

The analysis displayed here strongly advocates for public policy to be rigorously grounded in empirical and practical data. The responsibility of public officials extends beyond their required administrative duties; they are also tasked with fostering environments that lead to continuous improvement, wellbeing, and growth opportunities withing their communities. Understanding the effects that their public policies have on economic, demographic, and demographic indicators in their jurisdictions, will help them take more calculated and effective decisions, that could provide an improvement in key indicators of interests, like job growth and creation in this opportunity. Structured and informed governance is crucial for levering the advantages of new technologies, especially those helpful to improve the analysis necessary for improved public policy decision making.

8. Bibliography

Asamblea Nacional Constituyente. (1991). *Constitución Política de Colombia de 1991*. Imprenta Nacional.

Ascher, W. (1981). The forecasting potential of complex models. *Policy Sciences*, *13*(3), 247–267. https://doi.org/10.1007/BF00138485

Banco de la República. (2024). *Indice de Precio al Consumidor IPC Por Ciudades* [dataset]. https://totoro.banrep.gov.co/analytics/saw.dll?Portal

Behn, R. D. (2003). Why Measure Performance? Different Purposes Require Different Measures. *Public Administration Review*, *63*(5), 586–606. https://doi.org/10.1111/1540-6210.00322

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day. https://books.google.de/books?id=1WVHAAAAMAAJ

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brockwell, P. J., & Davis, R. A. (2010). *Introduction to time series and forecasting* (Second edition, (corrected at 8th. printing 2010)). Springer.

Congreso de la República de Colombia. (1994). *Ley 136 de 1994*.

DANE. (2019). *PROYECCIONES DE POBLACIÓN A NIVEL MUNICIPAL. PERIODO 2005—2019.* [dataset]. https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion

DANE. (2023a). *Gran Encuesta Integrada de Hogares* [dataset]. https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/mercado-laboral-historicos

DANE. (2023b). *Pobreza Monetaria* [dataset]. https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria

De Condorcet, M. (1795). Future progress of mankind. In *Outlines of an historical view of the progress of the human mind: Being a posthumous work of the late M. De Condorcet.* (pp. 316–372). J Johnson. https://doi.org/10.1037/11670-010

Departamento Nacional de Planeación. (2021). *Medición de Desempeño Municipal* [dataset]. https://portalterritorial.dnp.gov.co/AdmInfoTerritorial/MenuInfoTerrEstMDM

Departamento Nacional de Planeación. (2022a). *Historico Operaciones Efectivas de Caja* [dataset]. https://2022.dnp.gov.co/programas/desarrollo-territorial/Estudios-Territoriales/Informacion-Presupuestal/Ejecuciones-Presupuestales/Paginas/Operaciones%20Efectivas%20de%20Caja.aspx

Departamento Nacional de Planeación. (2022b). *Información fiscal y financiera* [dataset]. https://www.dnp.gov.co/LaEntidad_/subdireccion-general-descentralizacion-desarrollo-territorial/direccion-descentralizacion-fortalecimiento-fiscal/Paginas/informacion-fiscal-y-financiera.aspx

Dorado, D. (2021). *El Rol del Alcalde en el Desarrollo de los Mercados*.

Gerardo, C.-T., Robinson, C.-Á., & Asdrúbal, R.-M. (2014). *Impact of fiscal policy on savings and employment in colombia (1970-2011)*. *21*.

Goldsmith, S. (2017). *Análisis predictivo: Impulsar mejoras mediante el uso de datos*. https://blogs.iadb.org/administracion-publica/es/analisis-predictivo-impulsar-mejoras-mediante-uso-datos/

Hurwitz, J. (2018). *Machine Learning For Dummies®, IBM Limited Edition*.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

International Monetary Fund. (2013). Jobs and Growth—Analytical and Operational Considerations for the Fund. *Policy Papers*, *2013*(18). https://doi.org/10.5089/9781498342148.007

Jones, B. F., & Olken, B. A. (2005). *DO LEADERS MATTER? NATIONAL LEADERSHIP AND GROWTH SINCE WORLD WAR II*.

Martínez-Álvarez, J. J. (2015). *IMPACTO DE LAS REFORMAS ECONÓMICAS NEOLIBERALES EN COLOMBIA DESDE 1990. 8*(1).

Ministerio de Educación Nacional. (2024). *MEN ESTADISTICAS EN EDUCACION EN PREESCOLAR, BÁSICA Y MEDIA POR MUNICIPIO* [dataset]. https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/nudc-7mev/about_data

Ministerio de Salud y Protección Social. (2020, June 3). Colombia confirma su primer caso de COVID-19. *Colombia confirma su primer caso de COVID-19*. https://www.minsalud.gov.co/Paginas/Colombia-confirma-su-primer-caso-de-COVID-19.aspx#:~:text=Bogot%C3%A1%2C%206%20de%20marzo%20de,una%20paciente%20de%2019%20a%C3%B1os.

Ministerio del Trabajo. (2014). *Subdirección de Análisis, Monitoreo y Prospectiva laboral—Presentación institucional*.

OIT. (2013). *Modelo de Proyección de Empleo para Colombia*.

PNUD. (2016). *MERCADO LABORAL: PRODUCTIVIDAD Y COMPETITIVIDAD PARA EL DESARROLLO*. MERCADO LABORAL: PRODUCTIVIDAD Y COMPETITIVIDAD PARA EL DESARROLLO. https://www.undp.org/es/colombia/projects/mercado-laboral-productividad-y-competitividad

Presidencia de la República de Colombia. (2011). *DECRETO 4108 DE 2011*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=44622

Shearmur, R., & Polèse, M. (2007). Do Local Factors Explain Local Employment Growth? Evidence from Canada, 1971–2001. *Regional Studies*, *41*(4), 453–471. https://doi.org/10.1080/00343400600928269

Shumway, R. H., & Stoffer, D. S. (2016). *Time Series Analysis and Its Applications*.

Weber, M. (1977). *¿Qué es la burocracia?*

Wolber & Alfy. (1999). Monotonic cubic spline interpolation. *Proceedings Computer Graphics International CGI-99*, 188–195. https://doi.org/10.1109/CGI.1999.777953

Yan, X., & Su, X. (2009). *Linear regression analysis: Theory and computing*. World Scientific.

9. Statement of Authorship

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on April 29th of 2024 is identical to the printed version I submitted to the Examination Office on April 30th of 2024.

DATE: April 29th of 2024

NAME: Alvaro Jose Guijarro May

SIGNATURE: 

10. Statement of academic integrity related to the use of artificial intelligence tools

I acknowledge the use of artificial intelligence-based tools to support the writing of this thesis. This use was aimed at assistance on code writing in R language, translation from Spanish to English, and finding synonyms or alternative wordings to sentences.

DATE: April 29th, 2024

NAME: Alvaro Jose Guijarro May

SIGNATURE: 

11. Annex

Name of Variable	Source	Variables of Interest	Frequency & Time Frame	Description
Gran Encuesta Integrada de Hogares	(DANE, 2023a)	- **Workers** - Economic sector | Type of Economic Activity under CIIU 4 A.C - City | Name of a city in Colombia - Date | Year and Month	Monthly 2015-03/2023-12	Survey that contains the information of Colombian's employment conditions, in addition to general characteristics of the population such as sex, age, marital status and educational level, and asks about their sources of income. The GEIH provides the country with information at the national level, head, regional, departmental, and for each of the departmental capitals.
Population	(DANE, 2019)	- Population monthly | Total Population in monthly frequency (interpolated) - Population Yearly | Total Population in yearly frequency - City | Name of a city in Colombia - Month	Yearly 2010 - 2023	Population projections taking as base the 2018 Census methodology.
Consumer Price Index	(Banco de la República, 2024)	- Consumer Price Index (CPI) | Consumer Price Index, The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and medical care. The CPI is calculated by taking price changes for each item in the predetermined basket of goods and averaging them. - CPI year to date variation | % variance - The CPI (Consumer Price Index) year-to-date (YTD) variance refers to the change in the CPI from the beginning of the current year up to a specific point in time within the same year. - CPI yearly variation | % variance - The CPI (Consumer Price Index) yearly variance refers to the percentage change in the CPI over a 12-month period. It measures the rate of inflation or deflation by comparing the price level of the	Monthly 1973-01/2024-03	The consumer price index (CPI) measures the evolution of the average cost of a basket of goods and services representative of households' final consumption, expressed in relation to a base period. Calculated with data from DANE.

		CPI at the end of a year to the price level at the end of the previous year. - CPI monthly variation - City | Name of a city in Colombia - Date | Year and Month		
Education	(Ministerio de Educación Nacional, 2024)	- City | Name of a city in Colombia - Date | Year and Month - Year - Enrollment Rate 5-16 y.o | % - Proportion of the population between 5 and 16 years old who are attending the educational system. When DANE's population projections do not adequately capture internal migratory flows, it can reach values greater than 100%. - Net Coverage | % - It is the ratio between the number of students enrolled in transition, primary, secondary, and high school who have the theoretical age (5 to 16 years) and the total population of that same age. When DANE's population projections do not adequately capture internal migratory flows, it can reach values greater than 100% - Net Coverage Transition | % - It is the ratio between the number of students enrolled in transition who have the theoretical age to attend this level (5 years) and the total population of that same age. When DANE's population projections do not adequately capture internal migratory flows, it can reach values greater than 100%. - Net Coverage Primary | % - It is the ratio between the number of students enrolled in primary who have the theoretical age to attend this level (6 to 10 years) and the total population of that same age. When DANE's population projections do not adequately capture internal migratory flows, it can reach values greater than 100%. - Net Coverage Secondary | % - It is the ratio between the number of students enrolled in secondary who have the theoretical age to attend this level (11 to 14 years) and the total population of that same age. When DANE's population projections do not adequately capture internal migratory flows, it can reach values greater than 100%. - Net Coverage High School | % - # It is the ratio between the number of students enrolled in high school who have the theoretical age to attend this level (15 to 16 years) and the total population of that same age. When DANE's population projections do not adequately	Yearly 2011 - 2022	Contains statistical information on preschool, primary, secondary, and high school levels related to sector indicators by municipality without outliers, from 2011 to 2022.

		capture internal migratory flows, it can reach values greater than 100%. - Dropout Rate | % - # Intra-annual dropout rate of the official sector. Identifies the proportion of enrolled students who, due to cultural factors, conjunctural situations, or the provision of educational service, leave their studies during the academic year. - Dropout Rate Transition | % - Intra-annual dropout rate of the official sector in transition. Identifies the proportion of enrolled students who, due to cultural factors, conjunctural situations, or the provision of educational service, leave their studies during the academic year. - Dropout Rate Primary | % - Intra-annual dropout rate of the official sector in primary. Identifies the proportion of enrolled students who, due to cultural factors, conjunctural situations, or the provision of educational service, leave their studies during the academic year. - Dropout Rate Secondary | % - Intra-annual dropout rate of the official sector in secondary. Identifies the proportion of enrolled students who, due to cultural factors, conjunctural situations, or the provision of educational service, leave their studies during the academic year. - Dropout Rate High School | % - Intra-annual dropout rate of the official sector in high school. Identifies the proportion of enrolled students who, due to cultural factors, conjunctural situations, or the provision of educational service, leave their studies during the academic year. - Pass Rate | % - Pass rate of students in the official sector. Identifies the percentage of students in preschool, basic, and high school education who pass according to current educational plans and programs. - Pass Rate Transition | % - Pass rate of students in the official sector in transition. Identifies the percentage of students at this educational level who pass according to current educational plans and programs. - Pass Rate Primary | % - Pass rate of students in the official sector in primary. Identifies the percentage of students at this educational level who pass according to current educational plans and programs. - Pass Rate Secondary | % - Pass rate of students in the official sector in secondary. Identifies the		

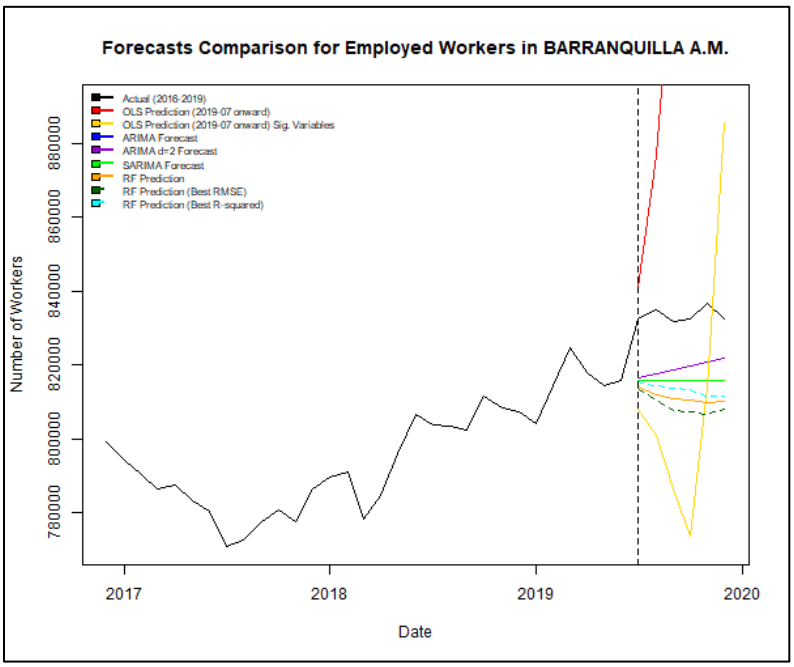
		percentage of students at this educational level who pass according to current educational plans and programs. - Pass Rate Highschool | % - Pass rate of students in the official sector in high school. Identifies the percentage of students at this educational level who pass according to current educational plans and programs. - Fail Rate | % - Failure rate of students in the official sector. Identifies the percentage of students in preschool, basic, and high school education who fail according to current educational plans and programs. - Fail Rate Transition | % - Failure rate of students in the official sector in transition. Identifies the percentage of students at this educational level who fail according to current educational plans and programs. - Fail Rate Primary | "% - Failure rate of students in the official sector in primary. Identifies the percentage of students at this educational level who fail according to current educational plans and programs. - Fail Rate Secondary | % - Repetition rate of the official sector. Corresponds to the percentage of students enrolled in secondary education who are repeating the same grade as the previous year. - Fail Rate High School | % - Repetition rate of the official sector. Corresponds to the percentage of students enrolled in high school who are repeating the same grade as the previous year.		
Monetary Poverty	(DANE, 2023b)	- I_PM | % of population - Monetary Poverty Rate - I_PME | % of population - Extreme Monetary Poverty Rate - Gini | Gini Coefficient (values between 0-1) - IPUG | $COP Values in Current Pesos - Average Per Capita Income of the Household Spending Unit - LP | $COP Values in Current Pesos - Monetary Poverty Lines (monthly values per person) - $COP Extreme Monetary Poverty Lines (monthly values per person), Values in Current Pesos - City | Name of a city in Colombia - Date | Year and Month	Yearly 2012-2022	Contains official monetary poverty figures of the Colombian population, corresponding to the methodological update based on information from the GEIH.
MDM Cities Indicators	(Departamento Nacional de	- MDM Resource Mobilization | Score between 1-100 - Measures mobilization of financial resources - Tax And Non-Tax Revenue Per Capita | $ COP Values in Current Pesos - Tax and non-tax	Yearly 2016 2022	Municipal Performance Measurement ("Medición de Desempeño

	Planeación, 2021)	revenue per capita, excluding territorial order collections - Revenue From OT Instruments Per Capita | $ COP Values in Current Pesos - Revenue collected through territorial ordering instruments per capita - Investment Financed By Own Resources | % - Percentage of investment financed by the municipality's own resources - MDM Execution Of Resources | Score between 1-100 - Execution of financial resources - MDM Open Government And Transparency | Score between 1-100 - Measures of open government and transparency practices - MDM Territorial Ordering | Score between 1-100, Territorial ordering and planning measures - Effective Collection Rate | Effective rate of tax collection - MDM Education | Score between 1-100 - Educational coverage and quality in middle education - MDM Health Coverage | Score between 1-100 - Health coverage and services - Health Coverage Overall | % of Population - Overall health coverage from the affiliate registry - Pentavalent Vaccination Coverage | % of Population - Coverage rate of the pentavalent vaccine in infants - Infant Mortality Rate | # of infant deaths - Infant mortality rate per 1000 live births - MDM Services | Score 1-100 - Coverage and quality of public services - Rural Electrical Coverage | % of Population - Coverage of rural electrical service - Broadband Penetration | % of Population - Number of broadband Internet subscribers relative to the total population - Aqueduct Coverage | % of Population - Coverage of aqueduct water service - Sewerage Coverage | % of Population - Coverage of sewerage service - MDM Security And Coexistence | Score 1-100 - Security and social coexistence indicators - Theft Rate Per 10k Inhabitants | # Reported theft cases per 10000 inhabitants - Homicide Rate Per 10k Inhabitants | # Homicide cases per 10000 inhabitants - Domestic Violence Rate Per 10k Inhabitants | # of Domestic violence cases per 10000 inhabitants		Municipal" MDM) aims to measure, compare, and rank municipalities according to their municipal performance, understood as management capacity and development results, taking into account their initial states.

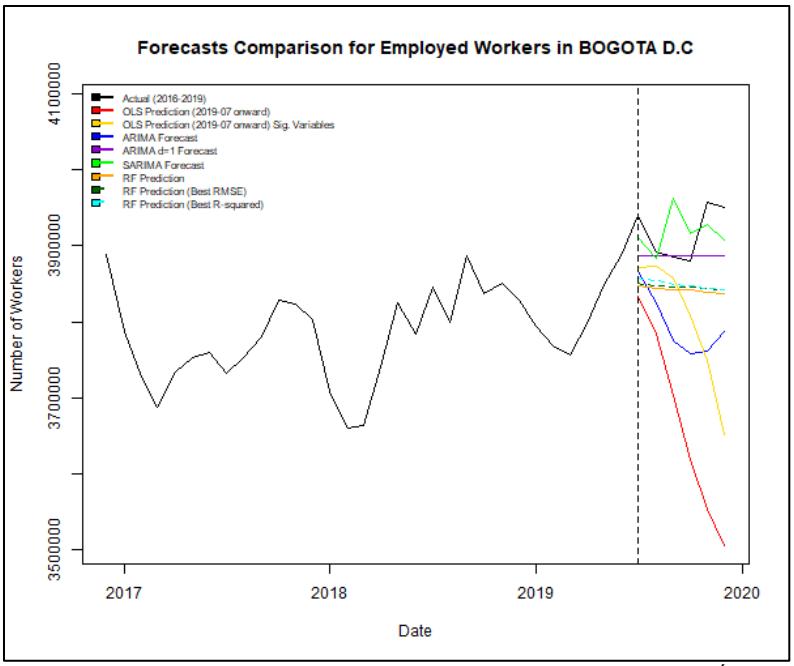
Fiscal Performance Amounts	Nacional de Planeación, 2022a)	- Total Income | $ Millions of Pesos - Total income received. - Current Income | $ Millions of Pesos - Current (or operational) income. - Tax Income | $ Millions of Pesos - Income received from taxes. - Property Tax | $ Millions of Pesos - Property tax income. - Industry And Commerce Tax | $ Millions of Pesos - Tax from industry and commerce activities. - Fuel Surcharge | $ Millions of Pesos - Surcharge on fuel. - Other Tax Income | $ Millions of Pesos - Other tax-related income. - Non-Tax Income | $ Millions of Pesos - Non-tax related income. - Current Transfers | $ Millions of Pesos - Current transfers received. - National Level Current Transfers | $ Millions of Pesos - Current transfers from the national level. - Other Transfers | $ Millions of Pesos - Other transfers. - Total Expenses | $ Millions of Pesos - Total expenses. - Current Expenses | $ Millions of Pesos - Current (or operational) expenses. - Operating Expenses | $ Millions of Pesos - Operating expenses. - Personal Services | $ Millions of Pesos - Expenses on personal services. - General Expenses | $ Millions of Pesos - General expenses. - Transfers Paid | $ Millions of Pesos - Transfers paid out. - Public Debt Interests | $ Millions of Pesos - Interests on public debt. - Current Dissaving Saving | $ Millions of Pesos - Current dissaving or saving. - Capital Income | $ Millions of Pesos - Income from capital. - Royalties | $ Millions of Pesos - Income from royalties. - National Transfers | $ Millions of Pesos - Transfers from the national level. - Co-financing | $ Millions of Pesos - Co-financing. - Other Capital Income | $ Millions of Pesos - Other capital income. - Capital Expenses | $ Millions of Pesos - Capital expenses.	Yearly 2000 2022	Municipal and departmental budget execution information aggregated in the Cash Flow Statement.

		- Gross Capital Formation | $ Millions of Pesos - Gross capital formation. - Other Capital Expenses | $ Millions of Pesos - Capital expenses. - Total Deficit Or Surplus | $ Millions of Pesos - Total deficit or surplus. - Financing | $ Millions of Pesos - Financing. - Net Credit | $ Millions of Pesos - Net credit. - Disbursements | $ Millions of Pesos - Net credit. - Amortizations | $ Millions of Pesos - Amortizations. - Balance Resources Variation In Deposits And Others | $ Millions of Pesos - Balance resources variation in deposits and others.		
Fiscal Performance Scores	(Departamento Nacional de Planeación, 2022b)	- Self-financing of operating expenses | Score 1-100 - Self-financing of operating expenses: the ability to cover the operating expenses of the central administration with unrestricted income (Law 617 of 2000) - Debt service support | Score 1-100 - Debt service support: the ability to support debt service with perceived revenues. - Dependence on transfers from the Nation and Royalties | Score 1-100 - Dependence on transfers from the Nation and Royalties: measures the importance of national transfers and royalties (SGR) in total revenues. - Generation of Own Resources | Score 1-100 - Generation of Own Resources: the ability to generate resources complementary to the transfers. - Magnitude of Investment | Score 1-100 - Magnitude of Investment: quantifies the magnitude of the investment executed by the territorial entity. - Saving Capacity | Score 1-100 - Saving Capacity: determines the degree to which surpluses are freed up to finance investment. - Fiscal Performance Indicator | Score 1-100 - Fiscal Performance Indicator - Category | Category - Type of Fiscal Performance of city	Yearly 2015 2022	Fiscal performance Score of the territorial entities for different fiscal years

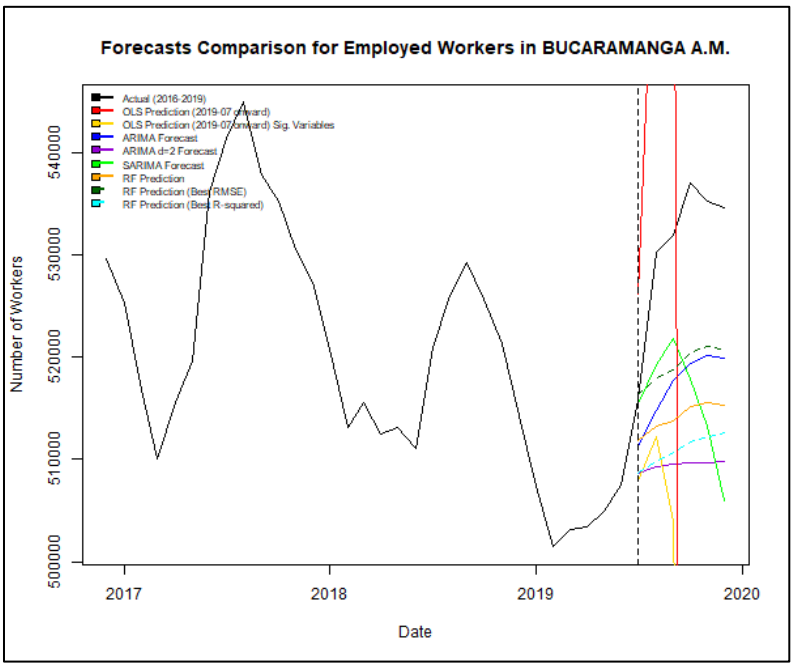
Annex 1 Description of variables gathered



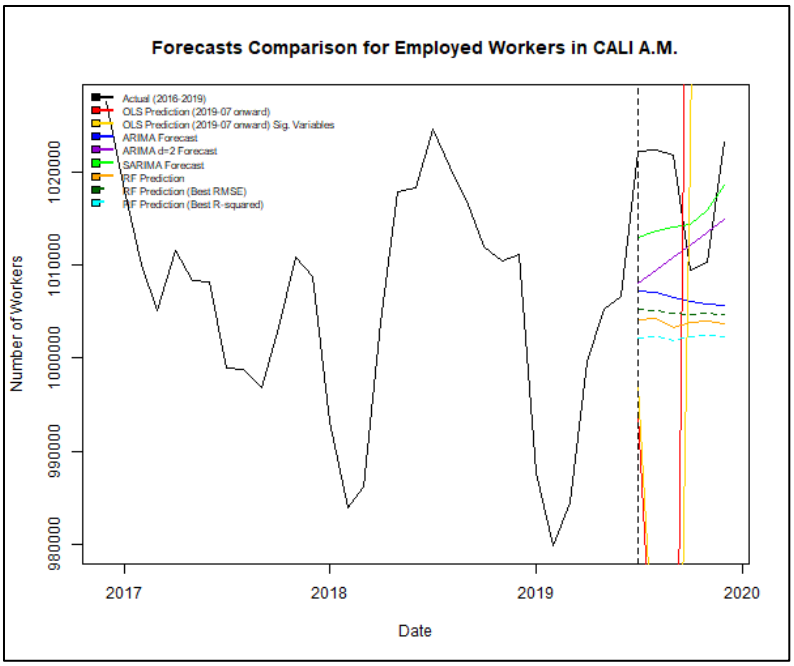
Annex 2, Prediction results for different model types for city of BARRANQUILLA A.M



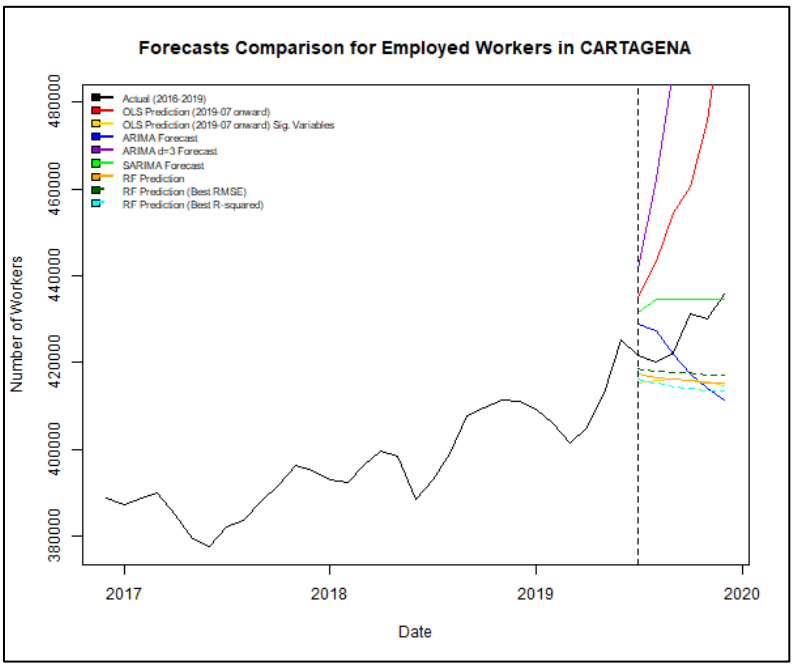
Annex 3, Prediction results for different model types for city of BOGOTÁ A.M



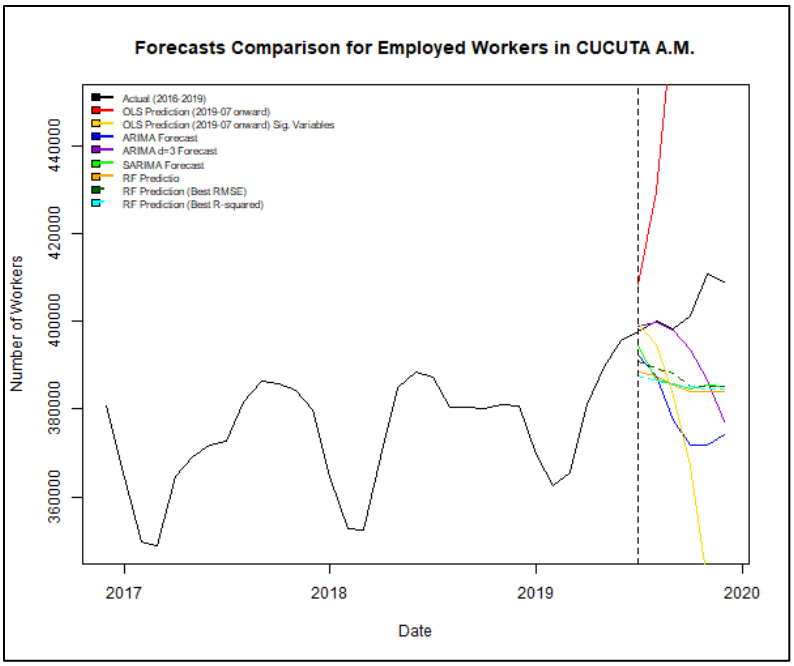
Annex 4, Prediction results for different model types for city of BUCARAMANGA A.M



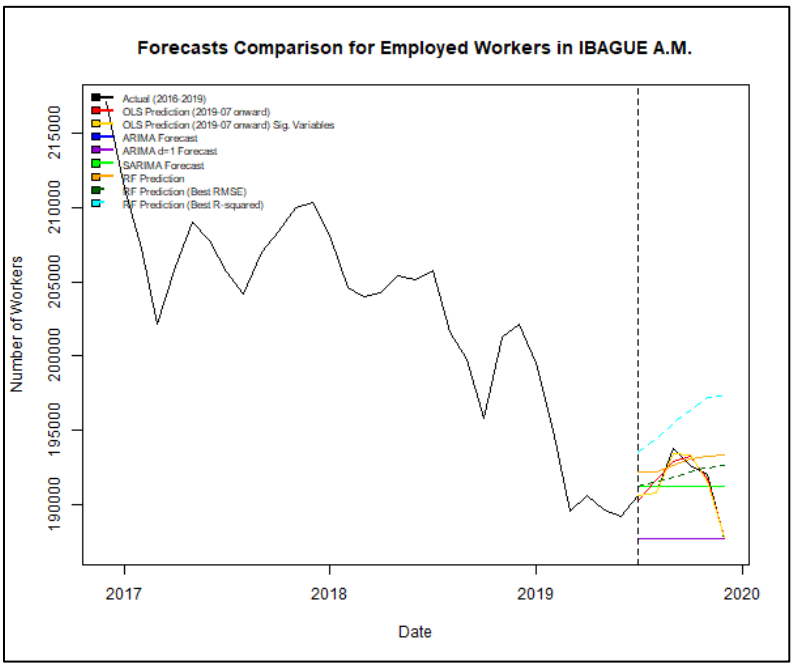
Annex 5, Prediction results for different model types for city of CALI A.M



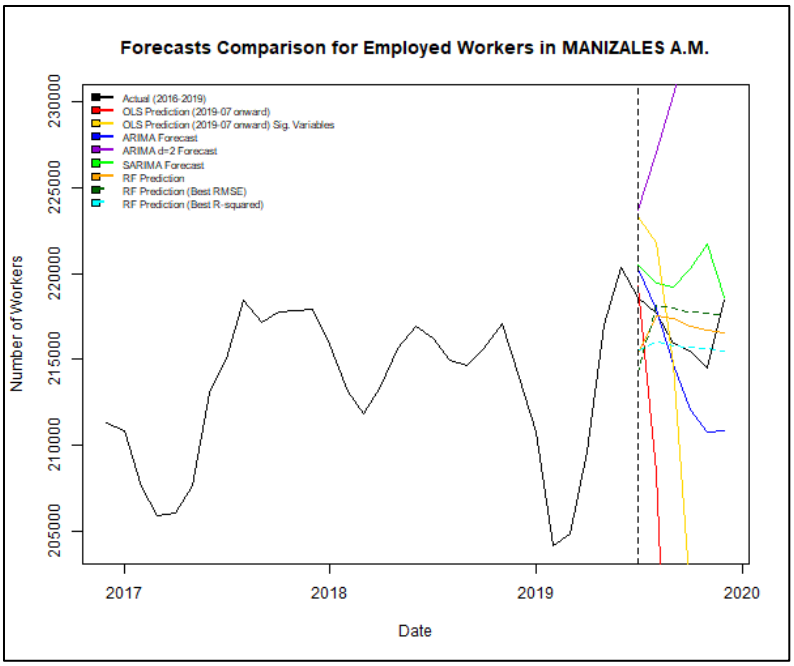
Annex 6, Prediction results for different model types for city of CARTAGEMA



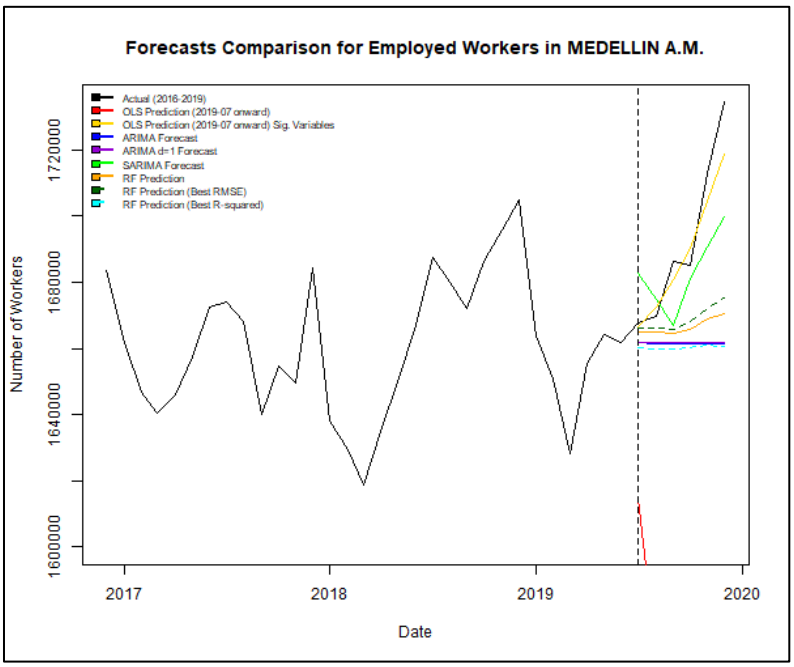
Annex 7, Prediction results for different model types for city of CUCUTA A.M



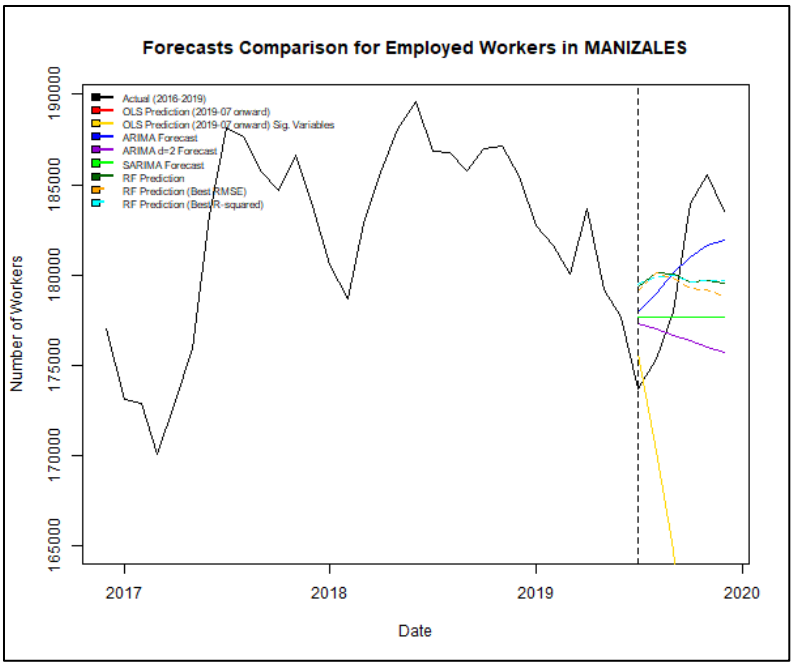
Annex 8, Prediction results for different model types for city of IBAGUE A.M



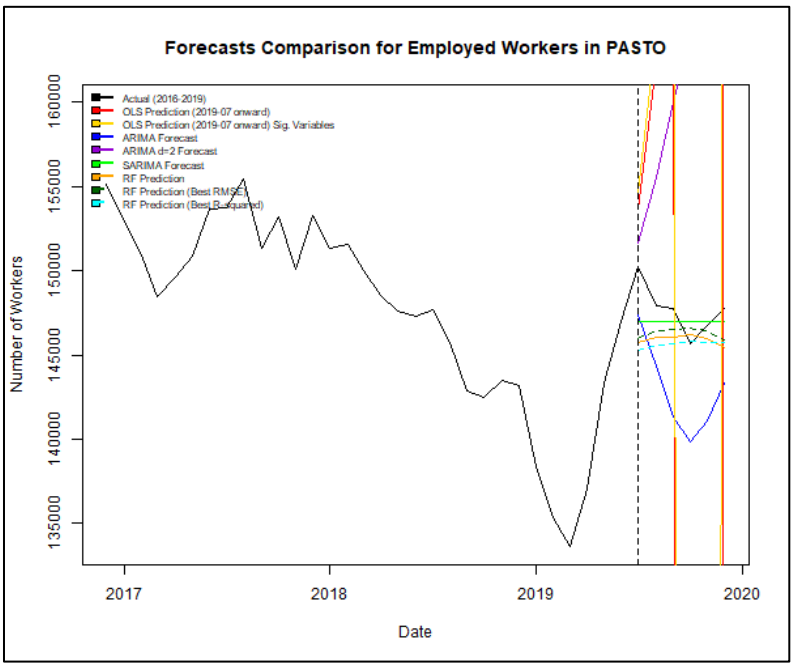
Annex 9, Prediction results for different model types for city of MANIZALES A.M



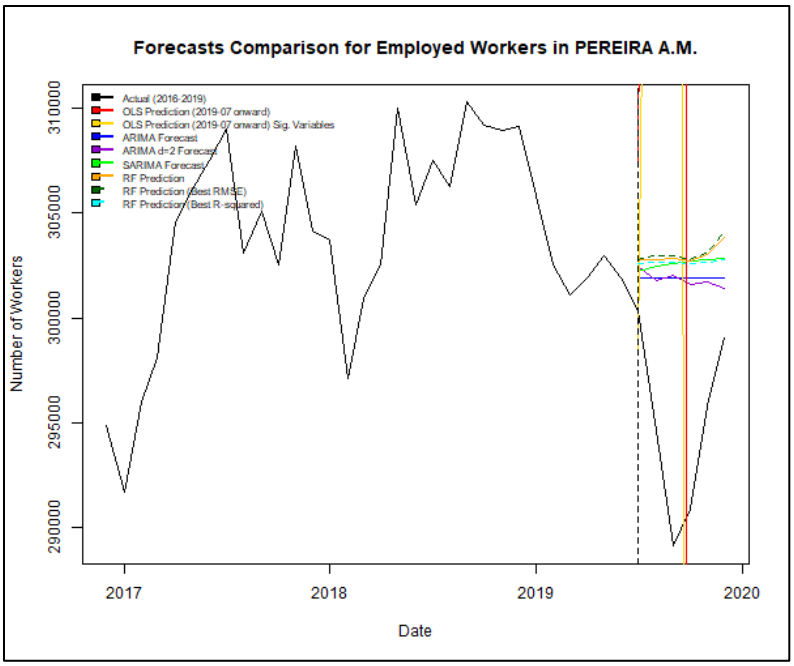
Annex 10, Prediction results for different model types for city of MEDELLIN A.M



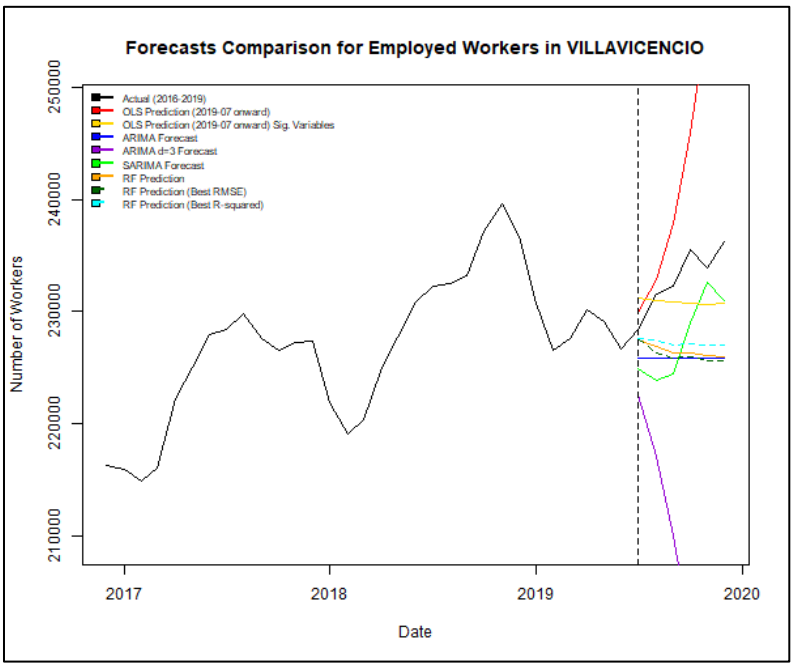
Annex 11, Prediction results for different model types for city of MANIZALES



Annex 12, Prediction results for different model types for city of PASTO



Annex 13, Prediction results for different model types for city of PEREIRA A.M



Annex 14, Prediction results for different model types for city of VILLAVICENCIO